

LA-UR-21-21459

Approved for public release; distribution is unlimited.

Title: Image Analysis: Comparison Metrics

Author(s): Aida, Toru

Intended for: Report

Issued: 2021-02-16

Disclaimer:

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Triad National Security, LLC for the National Nuclear Security Administration of U.S. Department of Energy under contract 89233218CNA000001. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Image Analysis: Comparison Metrics

Toru Aida

XCP-5, X Computational Physics Division, Los Alamos National
Laboratory

February 11, 2021

Introduction

Despite an increased reliance on computational modeling in engineering and physics, graphically comparing computational and experimental results has often been qualitative, via a so-called ‘viewgraph norm’ [1], [2]. Image recognition, computer vision and artificial intelligence (AI) are fields of study in themselves and rapidly progressing, but relying on AI to grade image similarity evokes a notion of asking for an expert judgement, which could be seen as an artificial version of the viewgraph norm. It is, therefore, desirable to use simpler metrics which are more tractable and unambiguous, even though they may not be as ‘intelligent.’

In this document, a few metrics are studied by systematically altering an image and their implications are analyzed.

Similarity Metrics

If A and B are $m \times n$ monochromatic image intensity maps, there are a few metrics that can be used as measures of their difference. These metrics are scalars so that the correlation between two 2-dimensional arrays (images) is summarized by single values. Probably the simplest is essentially the ‘size’ or the ‘absolute value’ of the difference of the two — mean squared error, root mean squared error or the Frobenius norm of the difference matrix [3]. They differ in terms of normalization or whether the square root is taken, but here the mean squared error (MSE) in Eq. 1 is used as their representative.

$$\text{MSE} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n (A_{ij} - B_{ij})^2. \quad (1)$$

An alternative, the Pearson product-moment correlation coefficient R between A and B , can be written as

$$R = \frac{\sum_{i=1}^m \sum_{j=1}^n (A_{ij} - \bar{A}) (B_{ij} - \bar{B})}{\sqrt{\left(\sum_{i=1}^m \sum_{j=1}^n (A_{ij} - \bar{A})^2 \right) \left(\sum_{i=1}^m \sum_{j=1}^n (B_{ij} - \bar{B})^2 \right)}} \quad (2)$$

where \bar{A} and \bar{B} are mean intensities of A and B , respectively [4]. $r = 1.0$ means A and B are identical, 0.0 means they are completely uncorrelated and -1 means that one is the negative of the other [5].

Structural Similarity Index Measure (SSIM) is another metric that was introduced by Wang et al. [6] which takes the form

$$\text{SSIM} = \frac{(2\mu_A\mu_B + c_1)(2\sigma_{AB} + c_2)}{(\mu_A^2 + \mu_B^2 + c_1)(\sigma_A^2 + \sigma_B^2 + c_2)} \quad (3)$$

where

$$\mu_A = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n A_{ij}, \quad \mu_B = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n B_{ij}, \quad (4)$$

$$\sigma_A = \sqrt{\frac{1}{mn-1} \sum_{i=1}^m \sum_{j=1}^n (A_{ij} - \mu_A)^2}, \quad \sigma_B = \sqrt{\frac{1}{mn-1} \sum_{i=1}^m \sum_{j=1}^n (B_{ij} - \mu_B)^2}, \quad (5)$$

$$\sigma_{AB} = \frac{1}{mn-1} \sum_{i=1}^m \sum_{j=1}^n (A_{ij} - \mu_A) (B_{ij} - \mu_B), \quad (6)$$

and $C_1 = (K_1L)^2$ and $C_2 = (K_1L)^2$ are small constants in order to avoid potential divide-by-zero errors, where L is the image intensity range. For this report the implementation in scikit-image [7] was used with the default parameters, meaning $K_1 = 0.01$ and $K_2 = 0.03$. The range of SSIM is -1.0 to 1.0, the same as R .

Image Correlation

As an example, a 750 pixel by 750 pixel, 8-bit grayscale image, meaning that the intensity ranges from 0 (black) to 255 (white), shown in Fig. 1, is designated as a target, or ‘experimental result.’



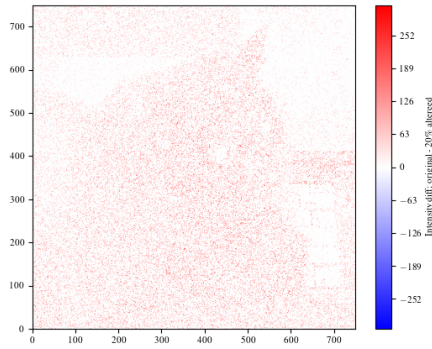
Figure 1: Original grayscale image.

This image was altered systematically to create ‘computational models,’ and the correlations between the ‘experiment’ and ‘model’ are presented below. First, the image was altered by replacing certain percentages of the pixels by random numbers ranging from 0 to 255¹. Figs. 2 show the difference, original – altered, meaning that red indicates the original is darker and blue, the original is lighter.

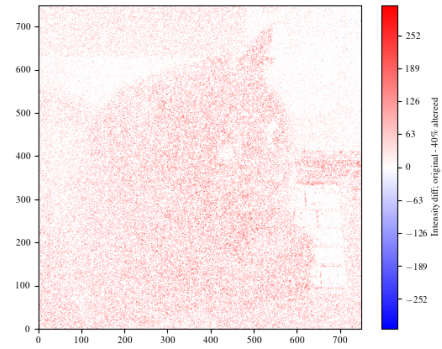
Note that the mean intensity of Fig. 1 is 115.9, meaning that on average it is on a darker side (< 128), and the random changes make the altered images lighter, rendering the difference images in Figs. 2 on the negative or red side. The average intensities after the alterations are 118.0, 119.7, 121.1 and 122.4 for 20%, 40%, 60%

¹Random changes were made using Python’s pseudo-random number generator (the Mersenne Twister) which has a reasonably long period, thus the randomness is fairly consistent.

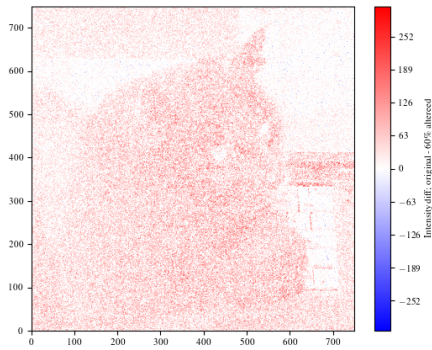
and 80% changes, respectively, and the average values of the difference images in Figs. 2 are -2.0, -3.8, -5.2 and -6.5, exemplifying the 'viewgraph-norm' where a cursory observation of these difference images can still reveal, at least qualitatively, the nature of alterations.



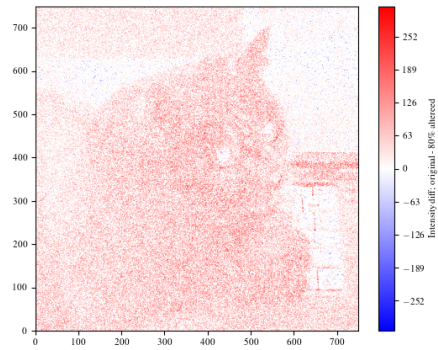
(a) 20% of the image is randomly altered.



(b) 40% of the image is randomly altered.



(c) 60% of the image is randomly altered.



(d) 80% of the image is randomly altered.

Figure 2: Intensity difference between the original (Fig. 1) and randomly altered versions.

R , SSIM and the inverse of MSE all correspond to the amount of changes that took place, as shown in in Figs. 3.

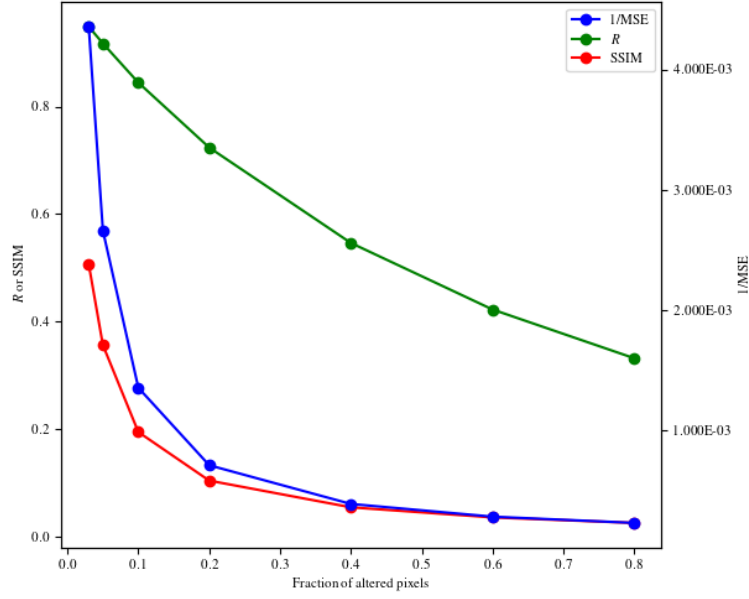
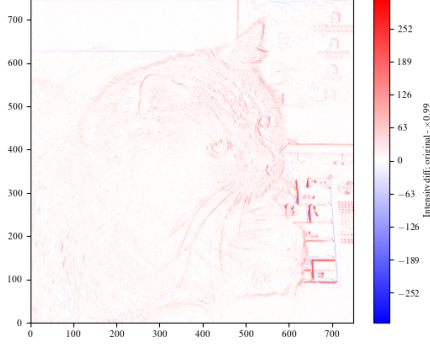


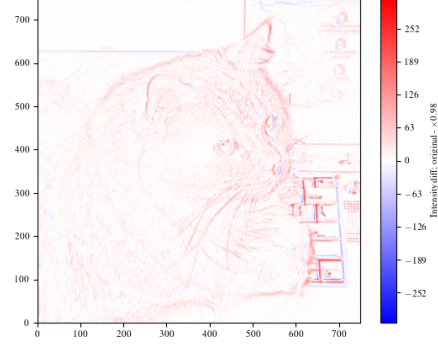
Figure 3: Pearson correlation coefficient R , SSIM and inverse of MSE, versus the percentages of pixels altered from Fig. 2.

Fig. 3 includes the metrics from images where 3%, 5% and 10% of the pixels are randomly altered (not shown in Figs. 2) to show how precipitously SSIM and $1/\text{MSE}$ drop with very minor alterations, and flatten as the alterations increase, meaning that once alterations to the image reach a certain level, further alterations do not result in significantly different SSIM or MSE. R , on the other hand, does not show such sensitivities.

Fig. 3 establishes that if the alterations are systematic, any of the metrics can be a fairly objective measure of the similarity between two images. This statement is true if the alterations are in the form of size-scaling or shifting by a few pixels (examples of scaled images shown in Figs. 4). The analysis of the metrics from these images will be elaborated later.



(a) Scaled by 99%.



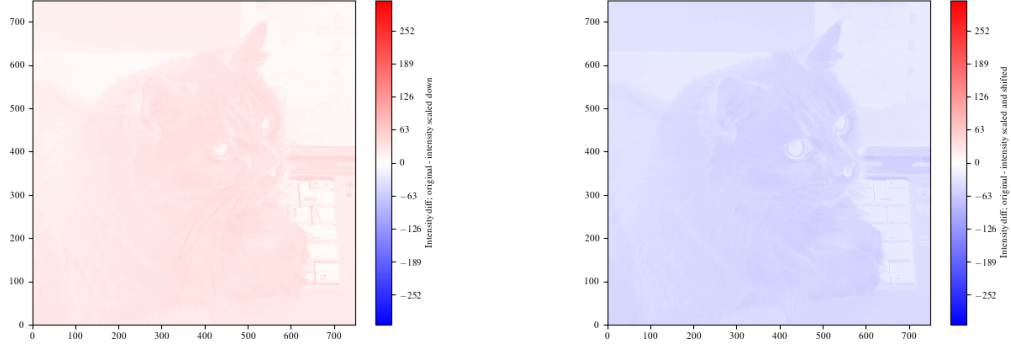
(b) Scaled by 98%.

Figure 4: Intensity difference the original (Fig. 1) and size scaled. $1/\text{MSE} = 4.932 \times 10^{-3}$, $R = 0.9532$ and $\text{SSIM} = 0.2713$ for 4a, $1/\text{MSE} = 3.355 \times 10^{-3}$, $R = 0.9310$ and $\text{SSIM} = 0.1067$ for 4b.

One of the issues, or the advantages, depending on applications, with the correlation coefficient R is that it is insensitive to linear transformation due to normalization with respect to the mean [5].

Figs. 5 show the differences when the original image intensity is scaled down by 20% ($\times 0.8$), and scaled up by 20% and shifted by -20 ($\times 1.2 - 20$) then compared against the original. They both have $R = 1.0$.

This holds true for the case where the original image intensity is shifted by 20 (not shown), or the case where the original image intensity is scaled up by 20% ($\times 1.2$; not shown).



(a) Intensity scaled down ($\times 0.8$).

(b) Intensity scaled ($\times 1.2$) and shifted (-20).

Figure 5: Intensity difference the original (Fig. 1) and intensity linearly scaled.

Another issue with the correlation coefficient R is that one for the whole image is not the average of those from subimages, as can be gleaned from Eq. 2. If an image with 20% of its pixels randomly altered is broken up in quadrants and metrics are calculated on each, the average SSIM and MSE of all the quadrants match those for the entire image. The average correlation coefficient R , however, does not (0.6512 for the average of all four quadrants whereas it is 0.7237 for the entire image in Fig. 6).

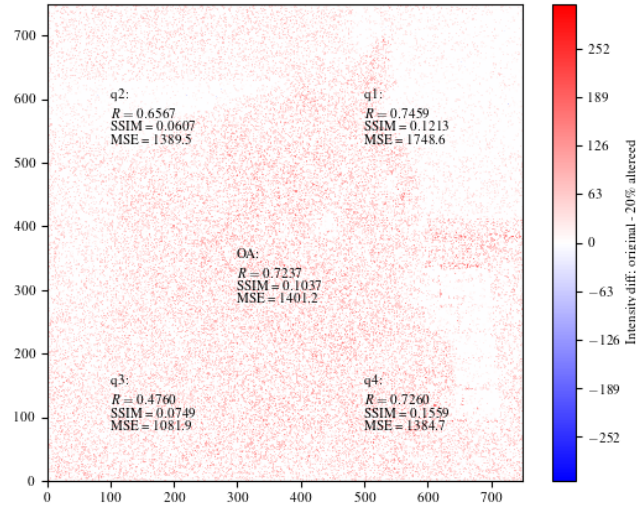


Figure 6: Similarity metrics; quadrant-wise vs. entire image.

Incidentally, the different metrics in different quadrants are due to the nature of the alterations and the characteristics of the original image; uniformly random changes cause the intensity histograms to flatten and increase the contrast. The intensity histogram of the third quadrant (Fig. 7b) is more ‘peaky’ than that of the first quadrant (Fig. 7a) and uniformly random changes cause more drastic shifts in the histogram.

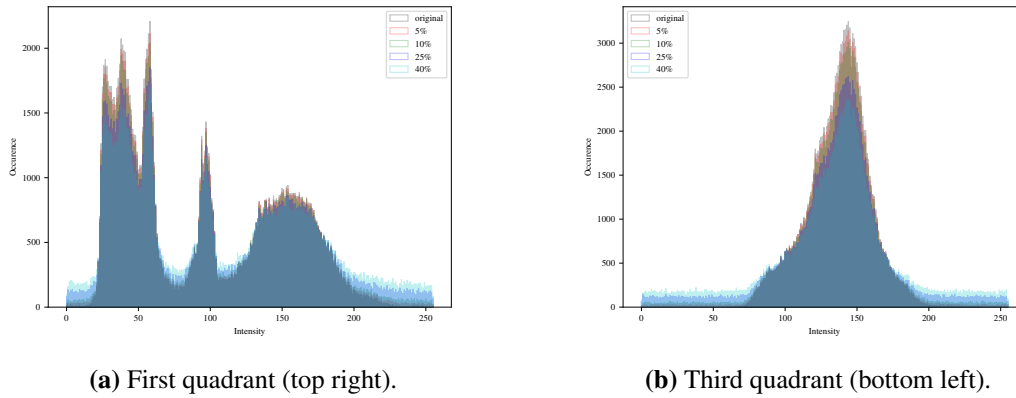


Figure 7: Intensity histogram changes due to the percentage of pixels randomly modified.

Discussions and Summary

Fig. 8 shows the summary of the difference metrics for all the altered images discussed above. Cases (a) through (d) correspond to a portion of Fig. 3.

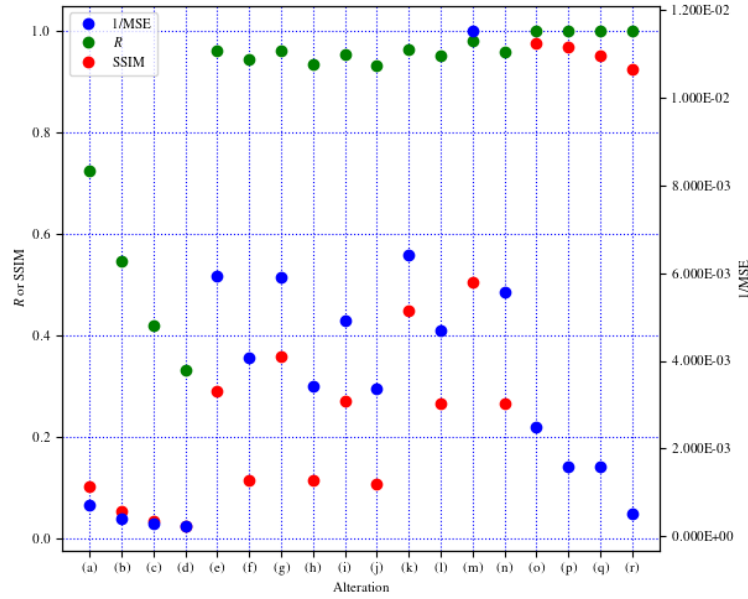


Figure 8: 1/MSE, R and SSIM for various alterations: (a) 20% random change, (b) 40% random change, (c) 60% random change, (d) 80% random change, (e) shifted 2 pixels to right, (f) shifted 4 pixels to right, (g) shifted 2 pixels down, (h) shifted 4 pixels down, (i) scaled down by 1%, (j) scaled down by 2%, (k) scaled down by 1% in x only, (l) scaled down by 2% in x only, (m) scaled down by 1% in y only, (n) scaled down by 2% in y only, (o) intensity shifted by 20, (p) intensity scaled up by 20%, (q) intensity scaled down by 20%, (r) intensity scaled up by 20% and shifted by 20.

As alluded to above, systematic changes result in predictable metrics; each of the pairs (e)-(f), (g)-(h), (i)-(j), (k)-(l), and (m)-(n) represents a small change in the former and a (relatively) large change in the latter of one form, and all the metrics for the former indicate a better match than the latter.

When different types of alterations are compared, there are some discrepancies among the metrics. For example, between 2 cases where the image is shifted by 2 pixels, case (e) (image shifted to the right) is deemed just as good (or bad) a match as (g) (image shifted down) based on R (0.9613 vs. 0.9611) or $1/MSE$ (5.926×10^{-3} vs. 5.914×10^{-3}) but SSIM clearly indicates that (e) is a worse match than (g) (0.2910 vs. 0.3598).

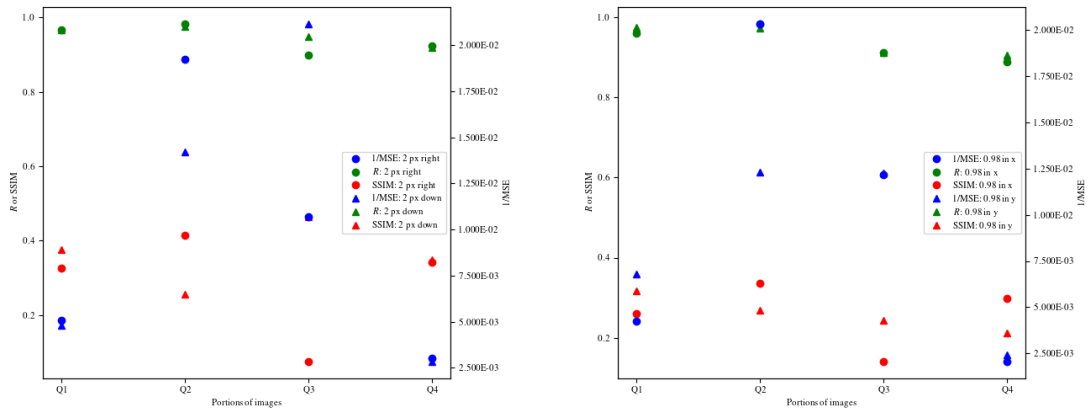
Between 2 cases where the image is scaled down by 2% in one direction only, R and

SSIM scores for (l) (only in x) and (n) (only in y) are similar (0.9511 vs. 0.9584 and 0.2652 vs. 0.2653, respectively), but 1/MSE indicates the latter is a better match than the former (4.697E-03 vs. 5.561E-03).

Cases (p) and (q) in Fig. 8 reveal another interesting comparison between the metrics. As mentioned above, these linearly transformed images resulted in $R = 1.0$, therefore R does not offer any insight as to the level of alterations. 1/MSE for them are the same, as the intensity of the both was scaled by the same amount (20%), but SSIM indicates that the scaled up version (p) is slightly better than the scaled down version (q).

These may be due to the fact that SSIM takes into account the changes in contrast, luminosity and the ‘structure’ of the image [6] while R and MSE are more or less simple comparisons in terms of luminosity (intensity) with no implied consideration for other features of the images. It is possible that the changes in different portions of the image cause one metric to be affected more than others, either compensating or accentuating the overall metrics.

For example, Fig. 9a shows that the difference in SSIM for the third quadrant between the shifted-to-the-right and shifted-down is much larger than that of MSE, thus SSIM indicating the shifted-to-the-right version is better, while the MSE scores are somewhat averaged out by those from the other quadrants. Similarly, the differences in the metrics for the forth quadrant in Fig. 9b cause SSIM to average out for the entire image, while the little discrepancy in MSE for that quadrant causes overall MSE to favor the one where the image is scaled in the y direction only, to be deemed better than the one where it is scaled in the x direction only. Note again that the averaging from different portions does not work for R .



(a) Shifted by 2 pixels; to the right by vs. down.

(b) Scaled by 98%; in x only vs. in y only.

Figure 9: Metric comparison by quadrants where overall metrics do not agree.

More striking example is one similar to what is presented in [5], where a blatant alteration is made whose average intensity is the same as the average intensity of the image. Table. 1 compares an image where 3% of the figure is randomly altered and one with a superimposed text, where all metrics imply that the latter is a much better match to the original than former.

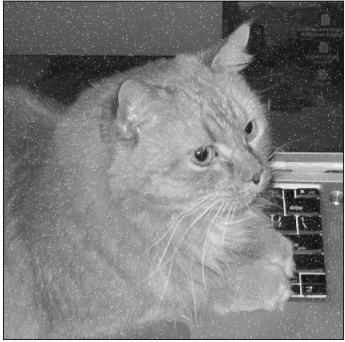
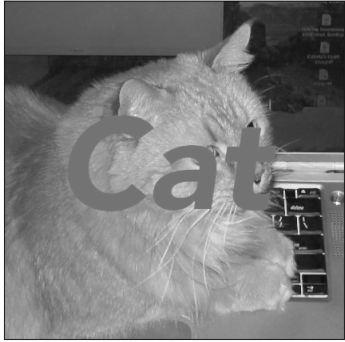
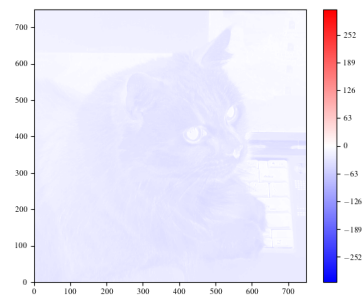
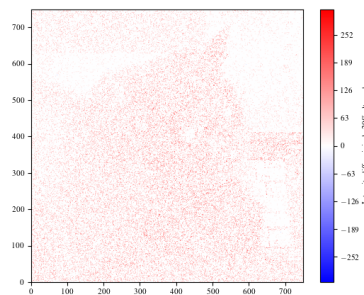
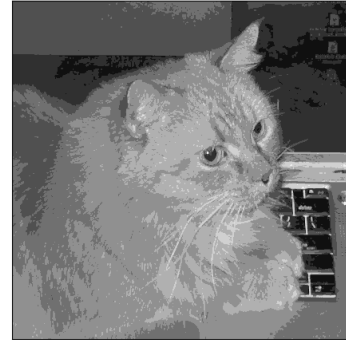
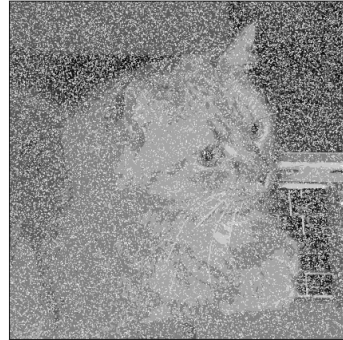
	3% of the pixels are altered.	Original with a superimposed text.
		
1/MSE	0.0043	0.0077
R	0.9487	0.9702
SSIM	0.5062	0.9009

Table 1: Image comparison where metrics do not correspond to the visual difference.

Furthermore, without obvious alterations, ‘better’ or ‘worse’ could be a matter of interpretation. For example, Table 2 shows 2 images where one is clearly considered better than the other in terms of all the metrics. However, it is obvious by the method of alteration that every pixel of the image on the right is ‘wrong,’ whereas 80% of pixels in the picture on the left match the original perfectly.

20% pixels altered (Fig. 2a).

$\times 1.2$ on all pixels².



1/MSE	7.138E-04	1.601E-03
<i>R</i>	0.7237	1.0000
SSIM	0.1037	0.9675

Table 2: ‘20%’ difference.

In conclusion, generally any of the 3 metrics examined above can be used as a measure of difference between images. However, it is prudent to check all to see if they agree, and when they do not, examine as to why, possibly checking them on various portions to see if those from one or more portions of the images are skewing the overall scores. As evidenced in the last example above, even when all metrics agree, blind reliance on them should not be automatic and an expert judgement is still relevant.

²By multiplying the intensity by 1.2, the range of pixel values actually changes to 0 – 306, making it technically not an 8-bit grayscale image. The metrics are calculated without converting it to 8-bit (i.e. not pruning the higher-than-255 values to 255).

References

- [1] W. L. Oberkampf, T. G. Trucano, and C. Hirsch, “Verification, Validation, and Predictive Capability in Computational Engineering and Physics,” Sandia National Laboratories, Albuquerque, NM, Tech. Rep., SAND2003-3769.
- [2] W. L. Oberkampf and M. F. Barone, “Measures of Agreement Between Computation and Experiment: Validation Metrics,” *Journal of Computational Physics*, vol. 217, pp. 5–36, Mar. 2006.
- [3] G. H. Golub and C. F. V. Loan, *Matrix Computations: Third Edition*. Baltimore and London: The Johns Hopkins University Press, 1996.
- [4] J. L. Rodgers and W. A. Nicewander, “Thirteen Ways to Look at the Correlation Coefficient,” *The American Statistician*, vol. 42, no. 1, pp. 59–66, Feb. 1988.
- [5] E. K. Yen and R. G. Johnson, “The Ineffectiveness of the Correlation Coefficient for Image Comparisons,” Los Alamos National Laboratory, Los Alamos, NM, Tech. Rep., LA-UR-96-2474.
- [6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image Quality Assessment: From Error Visibility to Structural Similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [7] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors, “Scikit-image: Image processing in Python,” *PeerJ*, vol. 2, e453, Jun. 2014, ISSN: 2167-8359. DOI: [10.7717/peerj.453](https://doi.org/10.7717/peerj.453). [Online]. Available: <https://doi.org/10.7717/peerj.453>.